Indirect Geospatial Referencing through Place Names in the Digital Library: Alexandria Digital Library Experience with Developing and Implementing Gazetteers

Linda L. Hill, Qi Zheng
University of California, Santa Barbara, California

Abstract

All types of information can be referenced to a geographic place. Maps, aerial photographs, and remote sensing images are spatially georeferenced. Other forms of information such as books, articles, research papers, pieces of music, and art are often linked to a geographic location through place names (geographic names). A gazetteer (a dictionary of geographic names) that is spatially referenced itself provides the bridge between these two types of georeferencing. With a georeferenced gazetteer translation service, a user can start with a geographic name and find information that is described with either geographic names or with geospatial coordinates. Use of this powerful indirect geospatially referencing tool can be applied as a common approach to libraries, bibliographic files, data centers, web resources, and museum and specimen collections and can be particular useful across language barriers since latitude and longitude coordinates are universally understood. The Alexandria Digital Library has implemented a gazetteer component for its georeferenced digital library. This experience resulted in the creation of a Gazetteer Content Standard, a Feature Type Thesaurus, and an operational interactive gazetteer service. This paper describes the development of these components and illustrates the use of this tool in a georeferenced digital library. It also relates progress in working with Federal agencies and others toward developing shareable gazetteer data through Digital Gazetteer Information Exchange programs.

INTRODUCTION

A gazetteer is a list of geographic names, together with their geographic locations and other descriptive information. A geographic name is a proper name for a geographic place or feature, such as *Washington, D.C.*, *Gulf of Mexico*, *Central High School*, and *Southern France*. Geographic places and features include political and administrative areas (e.g., cities, counties, and countries), natural features (e.g., mountain ranges, lakes, and canyons), manmade structures (e.g., buildings, bridges, and canals), and imprecise areas like *Southern California*. Gazetteer data exist in toponymic authority files (e.g., U.S. Board on Geographic Names gazetteers (U.S. Board on Geographic Names, 1998)), in published gazetteers (e.g., the New York Times Atlas (Mackay, John Bartholomew and Son, & Times Books, 1992)), in thesauri of geographic names (e.g., the Getty Thesaurus of Geographic Names (Getty Information Institute, 1997)), as biogeographic and physiographic regions (e.g., U.S. watersheds (Environmental Protection Agency, 1999)), as ecological regions (e.g., Bailey's Ecoregions (Bailey & Ropes, 1998)), as biological specimen collection sites and habitat areas, as census districts and topological map areas, and as countless administrative districts, study areas, and so forth. All *named geographic places* are proper features to be represented in a gazetteer.

A gazetteer in the context of an online information service (e.g., a digital library) is a knowledge base that, at a minimum, defines geographic names by spatial representations. With such relationships established between geographic names and their spatial footprints, two-way translations can take place: a geographic name can be translated into a spatial footprint and a spatial footprint can be translated into the set of geographic names in that area. When geographic names are also categorized by type (e.g., hydrographic features, populated places, etc.), specified types of places can be identified within a spatial area.

Beyond this minimum set of information for each place (name, footprint, and type), a gazetteer can

(1) provide variant names for the same location
(2) trace historical changes in names and in spatial footprints
(3) contain variant spatial representations for the same feature (a point location, bounding box, detailed boundary, etc. or spatial footprints from different sources) and measurement details
(4) link to (or include) other types of information about the feature (physical dimensions, history, description, population, etc.).
(5) link to related gazetteer entries, with links such as "Is Part Of" and "Is Capital Of"

(6) combine information from various sources together for one entry, with citation to the source of each contribution.

Figures 1 & 2 illustrate some of these elements of information with two examples of gazetteer entries from the current (6/99) ADL Gazetteer:

Gazetteer Record Report
Alexandria Digital Library

Feature Name
Geographic Name: *Santa Barbara Channel 1812-12-21 19:00 UTC Earthquake* (UCSB-INST CRUSTAL STUDIES-1)
Feature Type
ADL Feature Type Thesaurus: *earthquake features*
Spatial Reference
Geometry Type: *Point* (UCSB-INST CRUSTAL STUDIES-1)
Longitude: *-119.900002 (119°54'0"W)*
Latitude: *34.200001 (34°12'0"N)*
Measurement Method: *Earthquake Colocation*
Measurement Accuracy: *20km*
Identification Number
ADL Feature ID: *13001*
Feature Data
Seismic Moment Magnitude: *7.000000 M* (UCSB-INST CRUSTAL STUDIES-1)
Note: *The magnitude listed here is the "summary magnitude". For most events prior to 1898 this is the adjusted intensity magnitude, and for events after 1898 it is the surface wave magnitude.*
Related Information
Link to Related Information: *For more detailed information about this earthquake, see* http://www.crustal.ucsb.edu/ics/sb_eqs/1812/1812.html (UCSB-INST CRUSTAL STUDIES-1)
Note (UCSB-INST CRUSTAL STUDIES-1)
Earthquake Location Accuracy Assigned. Pre-1950: 20km, Post 1950: 2km.
Source
Source: *UCSB-INST CRUSTAL STUDIES-1*
Organization: *UCSB Crustal Studies*
URL: *http://www.crustal.ucsb.edu/*
Contact Person: *David Valentine*
Address: *UCSB, Santa Barbara, CA 93016, USA*
Information Source
Title: *California Earthquake History 1769-Present <http://www-socal.wr.usgs.gov/cahist_eqs.html>*
Author: *United States Geological Survey (USGS)*
Publisher: *United States Geological Survey (USGS)*
Publication Date: *March 31, 1998*

Figure 1. Example Record from the ADL Gazetteer

Gazetteer Record Report
Alexandria Digital Library

Feature Name
- Geographic Name: *Milano* (BGN-NIMA-1)
  - Language: *Italian*
  - Character Set: *ASCII*
- Variant Name: *Mailand* (BGN-NIMA-1)
- Variant Name: *Milan* (BGN-NIMA-1)
- Variant Name: *Mediolanum* (BGN-NIMA-1)

Feature Type
- ADL Feature Type Thesaurus: *populated places*
- NIMA Feature Designation: *PPL (populated place)*

Spatial Reference
- Geometry Type: *Point* (BGN-NIMA-1)
- Longitude: *9.200000 (9°11'59"E)*
- Latitude: *45.466667 (45°28'0"N)*

Identification Number
- ADL Feature ID: *2509048*

Related Information
- Related Feature (BGN-NIMA-1)
  - IsPartOf: *Italy*
  - IsPartOf: *JOG Sheet Number NL32-0*
  - IsPartOf: *Lombardia*
  - IsPartOf: *UTM grid NR13*

Source
- Source: *BGN-NIMA-1*
- Organization: *U.S. Board on Geographic Names (BGN), U.S. National Imagery and Mapping Agency (NIMA)*
- URL: *http://164.214.2.59/gns/html/index.html*
- Contact Person: *NIMA Geographer, D-56*
- Address: *4600 Sangamore Road, Bethesda, Maryland 20816-5003, USA*
- Information Source
  - Title: *Digital Interim Geographic Names Data (CDROM)*
  - Author: *U.S. National Imagery and Mapping Agency (NIMA)*
  - Series: *GAZGN, DIGNAMES*
  - Edition: *002*
  - Publisher: *U.S. National Imagery and Mapping Agency (NIMA)*
  - Publication Date: *August 1997*
  - Source ID: *NSN: 7644014174141*

Figure 2. Example Record from the ADL Gazetteer

A gazetteer can be viewed both as a freestanding reference service and as an embedded search aid in an information service. As a freestanding reference service (or collection), a user can ask "where-is" questions such as "Where is Niagara Falls"? or "What schools exist in an area and where are they located?" As an embedded search aid in an information service, a user may ask, "What remote-sensing images does the library have for the "Santa Barbara, California area"? Because the remote-sensing images that cover the Santa Barbara area are not specifically indexed to *Santa Barbara* as a place name, the query would first go to the gazetteer to obtain the geographic location which can then be used as a spatial query to find the relevant remote-sensing images. This use of gazetteer data is referred to as *indirect spatial referencing.*

The development of the gazetteer component of the Alexandria Digital Library has been previous described in a paper published in the electronic journal *D-Lib* (Hill, Frew, & Zheng, 1999). This paper will not include as much detail about the earlier process but will rather provide a summary of the lessons learned and more recent developments. Therefore, the reader is advised to read the previous paper for a complete picture of the project.

OVERVIEW OF THE ALEXANDRIA DIGITAL LIBRARY

The Alexandria Digital Library (ADL) <www.alexandria.ucsb.edu> is a distributed georeferenced digital library (a *geolibrary*) developed by a consortium of researchers, developers, and educators from academic, public, and private sectors and supported by an NSF-DARPA-NASA Digital Libraries Initiative, under cooperative agreement NSF IR94-11330. *Distributed* means the library's components may be spread across the Internet, as well as coexisting on a single desktop. A g*eoreferenced digital library* (a *geolibrary* (National Research Council Mapping Science Committee, 1999)) contains organized collections of objects where a primary attribute of those objects is their location on Earth in the form of a latitude and longitude point, a bounding box (two points of a box that includes the spatial extent of the object), a linear feature (e.g., path of a river), or a complete polygonal boundary. These coordinate representations of locations are called *footprints.*

The architecture of ADL is a 3-tier client-server architecture: client, middleware, and database servers (Frew et al., 1998). The ADL client is implemented in Java and designed to support interactive queries and evaluation of the search results. It includes an interactive map where the user can draw a *query area* as part of the query. Additional search parameters can be set using a set of high-level *search buckets* (Frew et al., 1999) designed to support single searches across multiple metadata sets in varying formats. The client supports enough local state to support the notion of sessions. The middleware performs the complex operations of communication and translation between multiple clients and multiple database servers holding the ADL collections. This involves creating and terminating logical sessions with clients, informing the clients about the collections that are available for searching, translating client queries to database queries, retrieving full metadata reports from the collections to pass on to the clients, and retrieving graphic representations (including reduced-resolution thumbnail views) of datasets or the actual data itself to pass on to the client and to the user for downloading.

An HTML client is also in development at the time of this writing. It is designed to give simpler and more easily accessible access to ADL. It is planned that this interface will be integrated into the California Digital Library <www.cdlib.org/> thus providing access to the georeferenced collections to a wide academic community.

Existing ADL collections include the following:

1) Catalog of aerial photographs, remote-sensing images, indexes to published map series, digital map products such as DEMs, DRGs, and DOQs, etc. provided by the Map & Imagery Laboratory (MIL) of the Davidson Library of the University of California, Santa Barbara
2) ADL Gazetteer
3) A subset of the GeoRef bibliographic database provided by the American Geological Institute (American Geological Institute, 1999)

The ADL homepage <www.alexandria.ucsb.edu> provides access to detailed descriptions of the collections and the types of items in the MIL Catalog. It also includes both a tutorial and a walkthrough that are useful for understanding the functionality of the system.

LESSONS LEARNED FROM IMPLEMENTATION OF THE ADL GAZETTEER

The initial ADL Gazetteer was built by combining the two major U.S. federal gazetteers into one. By far the hardest part of this process was combining the category schemes of the two sets into a coordinated set of types. This was accomplished by assigning the U.S. Geological Survey (USGS) categories to one of the nine feature classes from National Imagery and Mapping Agency (NIMA), creating a two-level hierarchy of types. Extensive analysis of the way the two type schemes had been implemented was necessary to create this common scheme. It served well for a place to start but also demonstrated that a hierarchical thesaurus of feature type terminology for gazetteers was needed to improve the structure of the categories.

For the initial gazetteer, we started with a minimal set of attributes for a gazetteer entry: Name, Location, and Type. We were aware, however, that information about places can be much more extensive than this and that it would be very beneficial if a content standard existed for gazetteer data to formalize the way in which this set of information was represented. Furthermore, if such a content standard could be widely adopted it would make the gathering and sharing of gazetteer data much easier.

The ADL Gazetteer proved to be a very valuable component of the digital library. The initial gazetteer held almost 6 million entries covering the United States and it territories (data from the USGS Geographic Names Service (U.S. Geological Survey, 1998)) and for the rest of the world (data from the NIMA GeoName Server (U.S. National Imagery and Mapping Agency, 1998)). In ADL, the Gazetteer is treated as a *collection* that can be selected and searched in the same way that the *catalog collection* (containing metadata for maps, aerial photographs, remote-sensing images and such) and other collections are treated. The Gazetteer is also an embedded search aid supporting indirect geospatial searching. The user can begin a search session by entering a place name, finding the right location, and then using that location to search the other collections spatially.

To support searching, the 6 million records in the Gazetteer had to be partitioned as 10 database files which were then searched in parallel to provide acceptable response times for the users. The search access included the capability to search by geographic name, type, and spatial area. Geographic name searching was supported by a degree of *fuzzy searching* to buffer the need to enter the names exactly as they were in the database. Filtering the retrieval by type (category) was supported by providing the user with the list of class and sub-classes used to categorize the places. Users could make a choice at the class level or at the sub-class level to limit the retrieval to places of a specified type or types. Geospatial searching was presented to the user as an interactive map that was used to specify the area of interest. A *spatial search datablade* of the underlying database was used to match the user's query area to the footprints of the gazetteer entries.

The initial experience of building a large gazetteer as part of the digital library lead to the next stages of development:

1) Developing a Gazetteer Content Standard
2) Developing a Feature Type Thesaurus
3) Implementing a new database structure
4) Reconverting gazetteer data sets to the new structure
5) Implementing new user interface and database access according to the new ADL architecture
6) Involving the federal government agencies in addressing the research and development issues of Digital Gazetteer Information Exchange (DGIE).

ADL GAZETTEER CONTENT STANDARD

A *content standard* specifies a common set of terminology and definitions for the documentation of data. It establishes the names of data elements and compound elements (groups of data elements) to be used for these purposes, the definitions of these compound elements and data elements, and information about the values that are to be provided for the data elements. (adapted from (U.S. Federal Geographic Data Committee, 1998)). The purpose of a content standard is to enable data sharing and distributed access through a common representation of information about data. Content standards guide the development of *metadata.*

ADL chose to develop a Gazetteer Content Standard (GCS) on the model of metadata rather than building or adopting a thesaurus structure for geographic names, the other obvious option, because a content standard supports

1) *contribution* of place descriptions to one gazetteer from multiple sources, and
2) *aggregation* of information from multiple gazetteers built in accordance with the Standard

A thesaurus, on the other hand, is not well suited to contribution and sharing between multiple originating sources. Additions to and modifications of a thesaurus go through an editorial body and must be carefully integrated with the existing structure and terms. Attempting to merge, or even map between, thesauri requires reconciling structural elements such as choice of hierarchical structure, preferred/non-preferred terms, and depth, specificity, and scope of coverage.

Development of the ADL Gazetteer Content Standard (GCS) was based on the experience of working with gazetteer data as well as a survey of the types of information contained in existing gazetteer data sets. The Standard attempts to be generic and flexible, with embedded documentation for the source of each piece of information and details about names, footprint measurements, and other elements. A link to the complete ADL GCS can be found at http://www.alexandria.ucsb.edu/gazetteer/.

Major sections of the Standard are ("(R)" indicates a Repeatable section):

1. Geographic Feature ID
2. Geographic Name
3. Variant Geographic Name (R)
4. Type of Geographic Feature (R)
5. Other Classification Terms (R)
6. Geographic Feature Code (R)
7. Spatial Location (R)
8. Street Address
9. Related Feature (R)
10. Description
11. Geographic Feature Data (R)
12. Link to Related Source of Information (R)
13. Supplemental Note
14. Metadata Information

Geographic Names and Variant Geographic Names can be described by (* indicates a Required element)

1. Name * [the primary name for feature in a particular the gazetteer application]
2. Name Source
3. Etymology
4. Language (default is English)
5. Pronunciation
6. Transliteration Scheme Used
7. Character Set (default is ASCII)
8. Current / Historical Note * (default is Current)
9. Beginning Date
10. Ending Date
11. Time Period Note
12. Source Mnemonic
13. Entry Date

Other sections of the Standard are similarly constructed with sub-elements. In addition, a *Source Content* Standard is specified to describe the organizations, individuals, and publications contributing the data.

It is a feature of a spatially-defined gazetteer that spatial hierarchy is inherently represented by the footprints (latitude and longitude coordinates). For example, the footprint of the city of Santa Barbara is within the footprint of Santa Barbara County, which is within the footprint of the state of California. Similarly, the city of Santa Barbara is also inherently part of particular physiographic, hydrographic, ecological, and mapping areas. It encompasses particular census districts. All of these relationships can be discovered from the spatial footprints. However, it is also desirable in some cases to explicitly declare the "Is Part Of" relationship. The GCS accommodates this in the "Related Feature" section. Some "Is Part Of" relationships are administrative and outside of spatial inclusion: Hawaii "Is Part Of" the United States administratively, for example. Correctly and fully representing inherent spatial relationships also requires "true" polygonal boundaries showing the extent of the places. In the ADL Gazetteer, and in most existing digital gazetteers, most of the footprints are points or bounding boxes rather than polygonal boundaries and thus do not do a good job showing true inclusion and overlapping spatial relationships.

An issue related to declaring the "Is Part Of" relationships is the format of the Geographic Name itself. In regular discourse we often identify a place by including a higher administrative entity: *Santa Barbara, California* or *Paris, France*. Part of the reason for this is to specify which Santa Barbara and which Paris we are talking about. In a spatially defined gazetteer, the footprints provide this specificity. Here even a point location (not a full polygonal boundary) is

sufficient to disambiguate *Paris, France* from *Paris, Texas*, for example. In the ADL Gazetteer, the Geographic Name values are the "simple" name of the place without added identification of a higher administrative body. If present in the record, this higher body will be recorded in the Related Feature section.

The CGS has been implemented by ADL and will continue to be modified, as necessary, by additional uses and other applications.

ADL FEATURE TYPE THESAURUS

To enable consistent description of types of places and features in the ADL Gazetteer, it was necessary to develop a new thesaurus of place/feature categories. The Feature Type Thesaurus (FTT) was done in accordance with the ANSI/NISO *Guidelines for the Construction, Format, and Management of Monolingual Thesauri* (Z39.19-1993) (National Information Standards Organization (U.S.) & American National Standards Institute, 1994). Candidate terms were obtained from the USGS Geographic Names Information Service (GNIS) and from the NIMA GeoNames Server as a beginning. Other gazetteer sets and geographic dictionaries were consulted for additional terms. For each set of gazetteer data converted to the ADL Gazetteer, its set of type terminology is integrated into the thesaurus. The result is a thesaurus that is very rich in *lead in* terms and that can therefore be used directly in the conversion process. Such extensive alternative terminology for single *concepts* also provides guidance for the search process. See the January 1999 *D-Lib* article for more details about the construction of the thesaurus.

A small set of top terms (major categories) was chosen to give structure to the hierarchy:

- Administrative Areas
- Hydrographic Features
- Land Parcels
- Manmade Features
- Physiographic Features
- Regions

The basic thesaurus design decisions were:

1) Generic relationships are implied between broader and narrower terms. That is, the narrower term is a member of the broader class.
2) Plural terms are used instead of singular terms.
3) Multiple hierarchies are allowed but not often used.
4) The depth of the hierarchy (i.e., the specificity of the preferred terminology) was heuristically determined based on the specificity likely to be needed by ADL. Terms more specific than needed by ADL were entered as non-preferred terms pointing to the broader term. If needed later, these non-preferred terms for more specific concepts can be changed to preferred narrower terms. For example, specific types of wetlands, such as *swamps* and *bogs*, are currently entered as non-preferred terms pointing to *wetlands*. A user looking for swamp features will be advised to use *wetlands* for the search. One reason for limiting the depth of the hierarchy is that ADL must perform feature type conversions when bringing in sets of gazetteer data and choice of feature type can be no more specific than can be justified by the clues in the incoming data.
5) All terms are in English. The MultiTes software that we are using, however, can handle parallel versions of thesauri in different languages and such an enhancement would be a natural development as the thesaurus matures.
6) Terms are not capitalized and are used in their natural order, rather than inverted. For example, *drainage basins* is used instead of *basins, drainage*.

Figure 3 shows an example of the contents of the ADL FTT. This is the entry for *wetlands* and shows the definition (Scope Note - SN), the non-preferred terms pointing to *wetlands* (Used For - UF), the Broader Term for *wetlands* (BT), and the Related Terms (RT) that are also preferred terms in the thesaurus. For other terms, Narrow Terms (NT) would also be shown.

A link to the complete thesaurus can be found at http://www.alexandria.ucsb.edu/gazetteer/.

wetlands
SN: A vegetated area that is inundated or saturated by surface or ground water for a significant part of the year. The vegetation is adapted for life in saturated soil conditions. [USGS 1048]
UF: backwaters
bayous
bogs
cienagas
fens
intermittent wetlands
mangrove swamps
marshes
mires
mud flats
peat cutting areas
peatlands
quagmires
salt marshes
sloughs
slues
swamps
tidal flats
BT: biogeographic regions
RT: bays
guts
lakes
playas
streams

Figure 3. Example of term entry from the ADL Feature Type Thesaurus

The relationship between a particular place and its type is an "Is A" relationship.

NEW DATABASE STRUCTURE

Based on the Gazetteer Content Standard, we designed a database, the ADL Gazetteer, using the Informix RDBMS. The database schema itself is not dependent on the Informix RDBMS. It only uses simple data types that are supported by almost all RDBMS systems such as Oracle, SQLserver, Sybase and Access. The entity relationship schema is a relatively straight translation of the Gazetteer Content Standard. The schema diagram is available as a PDF file at http://www.alexandria.ucsb.edu/gazetteer/. It will be adjusted as necessary as additional datasets are added. For presentation, some relationship links are omitted to reduce clustering on the diagram. Currently (6/99) the ADL Gazetteer database contains about 3,000,000 records from sources as NIMA, USGS, Southern California Earthquake Center, and others. We are in the process of adding 3,000,000 more U.S. features extracted from *the National Atlas of the United States* (U.S. Geological Survey, 1999) and GNIS.

CONVERSION ISSUES

In this section, we will describe our most recent conversion effort that involved the loading of the U.S. Geological Survey's Geological Name Information System (GNIS) data. The conversion issues for the NIMA data were described in the *D-Lib* article. These conversion issues are included to document the difficulty of sharing gazetteer data and in support of the Digital Gazetteer Information Exchange project described later in the paper.

Mapping the data elements of the GNIS record format to the ADL GCS was a straightforward process with little complication. The conversion of types, however, was very difficult. We also had to do special processing for GNIS records with multiple point locations.

Type Terminology

We have classified the categories of type conversion as the following:

- Specificity - where an incoming type is more specific than terms in the FTT
- Generality - where an incoming type is more inclusive than terms in the FTT
- Definition - where it isn't clear how to interpret the scope of an incoming category
- No category - where there has been no attempt to categorize the placenames

The NIMA type terms included many very specific categories that were converted to broader terms in the FTT. GNIS types, which are more general in scope, proved to be much more difficult to convert. They include several very broad categories (e.g., AREA, LOCALE, and OTHER). In these sets are types of places for which the FTT has specific terms. Categories other than the very broad ones (e.g., BAY) also contain a mix of types from the FTT. The assignment of FTT types to individual GNIS place names had to be based on the feature names themselves and the categories assigned by the GNIS. Explanatory notes are available occasionally but not often enough to make it possible to depend on them for semi-automatic conversion processes.

All through the GNIS conversion process, the FTT was modified by adding additional term variants and making changes to the structure and preferred terms as necessary. Any changes to structure were applied to the entries already in the new version of the ADL Gazetteer. This constant rethinking of the FTT as a result of incoming data is an important part of the process; it both validates and challenges the decisions that have been made and results in a more consistent and useful thesaurus.

Our process for converting the GNIS types to FTT terms was as follows.

1) First we processed each of the GNIS category sets separately so that we could take advantage of the GNIS type category to influence our type assignment.
2) A default FTT term was identified for each GNIS type. For the very general GNIS types (e.g., OTHER), no default could be designated, but for the others (e.g., BAY) the equivalent FTT term (e.g., *bays*) could be easily identified.
3) We created surname identification programs (perl scripts) that
    a) stripped a set of trailing words and numbers from the ends of the names
    b) identified frequently occurring *ending words* in names above a specified threshold (the *surnames*)
    c) checked the rest of the name for the occurrence of identified surnames in other positions of the name
    d) provided a total of the occurrence of the surnames as ending words and as keywords
    e) looked for two-word surnames in sets where there were many exceptions to regular one-word surname matching (e.g., "Shopping Center").
4) These scripts were used to examine each GNIS set. For exceptions to the default mapping for the set, individual surnames or keywords were mapped to FTT terms. This was done by direct lookup in the FTT if possible. If that was not possible, then thesaurus terms were supplied directly. Special case conversion (i.e., keywords and two-word surnames) were processed first; then single word surnames; and then the default FTT term was assigned to all remaining places in the GNIS set.

Figure 4 shows some examples of the surname mapping for places included in the GNIS AREA category.

Sample Surname and Keyword Mapping for the GNIS AREA Category

Surname	Keywords	Mapping to FTT
	Agricultural	agricultural sites
Beaver Dams		habitats
Camp Ground		camps
Demonstration Area		research areas
Fishery Area		fisheries
	Geothermal	thermal features
Glacier		glacier features
	Hunting	reserves
Indian Land		tribal lands
Mining Area		mine sites
	Petrified	petrified forests

Figure 4. Example of Surname and Keyword Mapping from GNIS Names to FTT Terms

The conversion programs are not 100% accurate in assigning FTT type terms to the GNIS records. Some of the sources of error are:

1) The surname is a false clue to the type of feature. Example, "Mitchell's Landing" as the name of a shopping center
2) The surname is not distinct enough to provide a correct clue to its type. Example, "The Rock"
3) Surname is misspelled. Example: "Birdge" instead of "Bridge"
4) Plural and singular forms of surnames.

Some of the complications of the conversions are:

1) In some cases, more than one FTT type needs to be assigned. Example: a name of the form " ... Bridge-Tunnel" needs to be assigned both *bridges* and *tunnels* as FTT types.
2) In some cases, the *keyword* that best represents the type of the place does not appear as a surname at the end of the name but rather appears at the beginning or within the name. Example: "Lake Powell" and "Eagle Island Prison Farm." But using keywords requires some care, as the last example illustrates. In this case, the conversion program treated "Prison" as a keyword to override "Farm" as a surname (and ignored the word "Island" in the name). For a name such as "Bayou Lake," the correct FTT type is *lakes* based on "Lake" as a surname; while a name such as "Bayou Pierre" should be given the type *wetlands* based on "Bayou" as a keyword. This requires some special programming logic based on the presence of recognizable surnames.
3) GNIS has more than 2 million records. Our surname extraction and matching programs were computationally expensive and time consuming for large subsets.

Multiple point locations

Features that extend across several topographic map sheets are given multiple points in the GNIS data. This applies to linear and large areal features. An example is the Mason-Dixon Line, which is represented in GNIS by a set of records each with multiple points. For the ADL Gazetteer, we needed to identify those cases with multiple points within one record and the duplicate records for one feature and combine them into one record. We also identified the maximum extent of the set of points for a single feature and created a *bounding box* footprint for the feature. This *bounding box* will be a generalization of the location of the feature. Although this will lead to some false hits in the search processing, it does allows ADL to represent the approximate location of the feature for search purposes.

ADL GAZETTEER SERVICE IMPLEMENTATION

A new development is the provision of a separate interactive Gazetteer Service to access the ADL Gazetteer. The ADL Gazetteer Server is accessible through http://www.alexandria.ucsb.edu/gazetteer/. The server has an HTML and Java based client interface and a backend supported by an AOL WebServer and Informix Universal RDBMS. The client contains a query form with an interactive map browser for location specification, a frame for the display of search results, and an area for viewing full gazetteer record information.

Using the ADL Gazetteer Server, users can form a query based on spatial location, feature type, words from place names, category terms provided by the originators of the data, and identifiers such as FIP codes and system control numbers. These five search parameters are part of ADL search buckets designed to support generic queries across multiple collections. The full ADL search bucket set also includes originators (not applicable to the Gazetteer), time periods (which we have not implemented yet for the gazetteer) and available formats (not applicable to the Gazetteer). It could be argued that the "originators" of the Gazetteer data are the source of the data (e.g., the USGS and NIMA), but this is at odds with the meaning of "originator" in the other ADL collections.

The ADL Gazetteer Server is supported by a gazetteer bucket database that is constructed by mapping and aggregating attributes in Gazetteer Content Standard to the five search buckets just described. The mapping is shown in Figure 5.

Search Bucket	GCS Attributes
Location	Bounding Box Coordinates
Type	ADL Feature Types (FTT Terms)
Free Text	Geographic Names, Variant Names, and Feature Description
Assigned Terms	Classification terms provided with the data (e.g., NIMA categories)
Identifiers	Feature Codes and ADL feature IDs

Figure 5. ADL Search Bucket Mapping for the GCS

The information and database schema diagram for the gazetteer buckets can be found at http://www.alexandria.ucsb.edu/gazetteer/.

It is easy to see from the schema diagram that it is very much dependent on the special features provided by the Informix Universal Server and its datablades: the MapInfo Spatial DataBlade and Verity Text Search DataBlade (VTS). The MapInfo Spatial DataBlade provides all of the spatial search operations. The VTS provides text-searching capability. For text searching, we have implemented the options of "All words" (AND operation), "Any words" (OR operation), and "Exact phrase," and stemming to provide the necessary *fuzzy* searching for place names.

The gazetteer server is currently running on a SUN E-3000 server. The majority of queries, against a database of over 3 million entries, have a response time of less than 10 seconds. This speed is partly due to the power of the SUN server but also to performance tuning on the server and on the SQL query processing.

DIGITAL GAZETTEER INFORMATION EXCHANGE PROGRAM

A meeting was held at the U.S. Geological Survey on December 11, 1998 to discuss gazetteer information exchange among federal government agencies. There were 22 attendees from 8 government agencies (U.S. Geological Survey, National Imagery and Mapping Agency, National Aeronautics and Space Agency, National Oceanic and Atmospheric Administration, Census, National Park Service, Library of Congress, & Smithsonian), one professional association (American Geological Institute), and one university (University of California, Santa Barbara). The group agreed to call itself the Digital Gazetteer Information Exchange (DGIE) planning group and to discuss proposals to advance the development of shareable gazetteer data in support of government information services.

The decision was made to apply for a grant from the National Science Foundation as part of their Digital Government Program to hold a DGIE workshop in the fall in Washington, DC. UCSB received the award. Linda Hill is the Principal Investigator (PI) and Michael Goodchild is the co-PI. It will be held on October 12-14 at the Smithsonian Institute's International Center. The two-day workshop will address the technical and policy challenges of building the means by which gazetteer data may be more readily integrated into web-based information services (digital libraries and information centers) and shared among the various producers and users of gazetteer data. The web page for the workshop is at http://www.alexandria.ucsb.edu/~lhill/dgie/DGIEworkshop.htm.

Workshop participants, limited to no more than 60, will be selected through a combination of invitation and open call. We anticipate participation from government agencies, academia, and the private sector; representation from overseas and representation of diverse cultural and under-represented groups, to include:

- gazetteer producers
- digital library developers
- librarians
- geospatial data center staff
- georeferenced clearinghouse and information infrastructure staff
- natural history museum staff
- researchers whose activities create or use georeferenced place/feature data (e.g., biodiversity, geography, ecology, geology, epidemiology, natural history, culture)
- computer scientists
- information science researchers

The expected benefits of the workshop include

1. Clarification of the concept of spatially-defined gazetteers and their relationship to information services (in particular, digital government information dissemination).
2. Facilitation of greater collaboration and synergy between various participants in this area - an interest area not previously identified as such.
3. Increased awareness of the potential of spatially-defined gazetteers among a much wider audience than the current limited gazetteer community, including digital libraries and digital Earth information systems.
4. Conceptual overview of Digital Gazetteer Information Systems/Services, identifying the requirements for software development, standards and protocols, data, and basic research.
5. A major proposal for research under the Digital Government Initiative program that will address substantive needs of agencies, partner with one or more agencies, increase access to Federal data, and provide basic research results in the areas of intelligent information integration and cost-effective acquisition, integration, viewing, and using large sets of spatially referenced place and feature name data sets.
6. Stimulus to researchers from computer and information science to work on problems associated with gazetteers.

CONTINUING DEVELOPMENTS

When the GNIS data is completely loaded into the ADL Gazetteer, the Gazetteer will contain more than 6 million entries from more than five sources. Other sets are waiting to be processed, including place names from the *GeoRef Thesaurus* and place names extracted from the USGS *National Atlas of the United States*. We hope to encourage holders of gazetteer data to convert their own information to the ADL Gazetteer Content Standard and to map their categories to the ADL Feature Type Thesaurus so that merging data is easier to do. An even better outcome would be for other gazetteer services to be built based on our work so that we can begin to build a network of gazetteer services.

Our development plans include creating gazetteer ingest software, with an interactive map; building the capability to process and display geographic names with extended ASCII character sets; and enhancing the entries with more polygonal boundaries and/or bounding boxes, more historical places and temporal ranges, and more extensive descriptions.

The outcome of the Digital Gazetteer Information Exchange workshop should also identify the goals and the related issues for enabling gazetteers in networked information systems in support of indirect geospatial referencing.

ACKNOWLEDGEMENTS

The ADL Gazetteer project has been supported by funding from NSF, DARPA, and NASA under NSF IR94-11330 and the NASA Earth Science Data and Information System. Graduate students Doug Tallman and Ajay Kochhar have provided programming assistance. We would also like to thank the members of the ADL Implementation Team and the Map and Imagery Laboratory of the Davidson Library at the University of California, Santa Barbara.

REFERENCES

American Geological Institute. (1999). *GeoRef*. http://www.agiweb.org/agi/georef/about.html.

Bailey, R. G., & Ropes, L. (1998). *Ecoregions : the Ecosystem Geography of the Oceans and Continents : with 106 illustrations, with 55 in color*. New York: Springer.

Environmental Protection Agency. (1999). *Locate Your Watershed Webpage*. http://www.epa.gov/surf2/locate/.

Frew, J., Freeston, M., Freitas, N., Hill, L., Janee, G., Lovette, K., Nideffer, R., Smith, T., & Zheng, Q. (1998). The Alexandria Digital Library architecture. In C. Nikolaou & C. Stephanidis (Eds.), *Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries (ECDL'98), Heraklion, Crete, Greece, Sept. 1998* (pp. 61-73). Berlin: Springer-Verlag. http://www.springer.de/comp/lncs/index.html.

Frew, J., Freeston, M., Hill, L., Janee, G., Larsgaard, M., & Zheng, Q. (1999). Generic query metadata for geospatial digital libraries. *Proceedings of the Third IEEE Meta-Data Conference (Meta-Data '99), April 6-7, 1999, Bethesda, MD, sponsored by IEEE, NOAA, Raytheon ITSS Corp., and NIMA*. . http://computer.org/conferen/proceed/meta/1999/papers/55/jfrew.htm.

Getty Information Institute. (1997). *Thesaurus of Geographic Names.* http://www.ahip.getty.edu/tgn_browser/.

Hill, L. L., Frew, J., & Zheng, Q. (1999). Geographic names: The implementation of a gazetteer in a georeferenced digital library. *D-Lib* (January 1999). http://www.dlib.org/dlib/.

Mackay, D., John Bartholomew and Son, & Times Books. (1992). *The New York Times Atlas of the World*. (3rd rev. concise, 3rd US ed.). New York, N.Y.: Times Books.

National Information Standards Organization (U.S.) & American National Standards Institute. (1994). *Guidelines for the Construction, Format, and Management of Monolingual Thesauri : an American National Standard*. Bethesda, Md.: NISO Press.

National Research Council Mapping Science Committee. (1999). *Distributed Geolibraries : Spatial Information Resources. Summary of a Workshop held June 15-16, 1998.* Washington, DC: National Academy Press.

U.S. Board on Geographic Names. (1998). http://www-nmd.usgs.gov/www/gnis/bgn.html and http://164.214.2.59/gns/html/BGN.html.

U.S. Federal Geographic Data Committee. (1998). *Content Standard for Digital Geospatial Metadata.* http://fgdc.er.usgs.gov/metadata/contstan.html.

U.S. Geological Survey. (1998). *Geographic Names Information Service (GNIS).* http://mapping.usgs.gov/www/gnis/.

U.S. Geological Survey. (1999). *National Atlas of the United States.* http://www-atlas.usgs.gov.

U.S. National Imagery and Mapping Agency. (1998). *Geonet Names Server.* http://164.214.2.59/gns/html/index.html.